NeuroImage 115 (2015) 256–268

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

# Fast and powerful heritability inference for family-based neuroimaging studies

Habib Ganjgahi [a], Anderson M. Winkler [b,c], David C. Glahn [c,d], John Blangero [e], Peter Kochunov [f], Thomas E. Nichols [a,b,g,*]

[a] Department of Statistics, The University of Warwick, Coventry, UK
[b] Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK
[c] Department of Psychiatry, Yale University School of Medicine, New Haven, USA
[d] Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, Hartford, CT, USA
[e] Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
[f] Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA
[g] WMG, The University of Warwick, Coventry, UK

ARTICLE INFO

*Article history:*
Received 19 December 2014
Accepted 3 March 2015
Available online 23 March 2015

*Keywords:*
Heritability
Permutation test
Multiple testing problem

ABSTRACT

Heritability estimation has become an important tool for imaging genetics studies. The large number of voxel- and vertex-wise measurements in imaging genetics studies presents a challenge both in terms of computational intensity and the need to account for elevated false positive risk because of the multiple testing problem. There is a gap in existing tools, as standard neuroimaging software cannot estimate heritability, and yet standard quantitative genetics tools cannot provide essential neuroimaging inferences, like family-wise error corrected voxel-wise or cluster-wiseP-values. Moreover, available heritability tools rely on P-values that can be inaccurate with usual parametric inference methods.

In this work we develop fast estimation and inference procedures for voxel-wise heritability, drawing on recent methodological results that simplify heritability likelihood computations (Blangero et al., 2013). We review the family of score and Wald tests and propose novel inference methods based on explained sum of squares of an auxiliary linear model. To address problems with inaccuracies with the standard results used to find P-values, we propose four different permutation schemes to allow semi-parametric inference (parametric likelihood-based estimation, non-parametric sampling distribution). In total, we evaluate 5 different significance tests for heritability, with either asymptotic parametric or permutation-basedP-value computations. We identify a number of tests that are both computationally efficient and powerful, making them ideal candidates for heritability studies in the massive data setting. We illustrate our method on fractional anisotropy measures in 859 subjects from the Genetics of Brain Structure study.



## Introduction

Combining neuroimaging data with genetic analyses is an increasingly active area of research aimed at improving our understanding of the genetic and environmental control over brain structure and function in health and illness (see, e.g., Glahn et al., 2007). The foundation of any genetic analysis is establishing that a trait is heritable, that is, that a substantial fraction of its variability can be explained by genetic factors. Significant and reproducible heritability has been established for many neuroimaging traits assessing brain structure and function, including, for instance, location and strength of task-related brain activation (Blokland et al., 2008; Koten et al., 2009; Matthews et al., 2007; Polk et al., 2007), white matter integrity (Kochunov et al., 2014a, b; Jahanshad et al., 2013; Brouwer et al., 2010; Chiang et al., 2009, 2011; Kochunov et al., 2010), cortical and subcortical volumes, cortical thickness and density (Winkler et al., 2010; Rimol et al., 2010; Kochunov et al., 2011a, b; Kremen et al., 2010; den Braber et al., 2013).

Variance component models are the best-practice approach for deriving heritability estimates based on familial data (Almasy and Blangero, 1998; Blangero and Almasy, 1997; Amos, 1994; Hopper and Mathews, 1982), for allowing great flexibility in modeling of genetic additive and dominance effects, as well as common and unique environmental influences. The model also allows the inclusion of additional terms that allow linkage analysis, yet remaining relatively simple and requiring the estimation of only a few parameters. Estimation of parameters typically uses maximum likelihood under the assumption that the additive error follows a multivariate normal distribution. The iterative optimization of the likelihood function requires computationally

* Corresponding author at: Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. Fax + 44 24765 24532.
*E-mail address:* t.e.nichols@warwick.ac.uk (T.E. Nichols).

http://dx.doi.org/10.1016/j.neuroimage.2015.03.005
1053-8119/

intensive procedures, that are prone to convergence failures, something particularly problematic when fitting data at every voxel/element.

Typically a likelihood ratio test (LRT) is used for heritability hypothesis testing. As the null hypothesis value is on the boundary of the parameter space, the asymptotic distribution of LRT is not $\chi^2$ with 1 degree of freedom (DF), but rather approximately as a 50:50 mixture of $\chi^2$ distributions with 1 and 0 DF, where a 0 DF $\chi^2$ is a point mass at 0 (Chernoff, 1954; Self and Liang, 1987; Stram and Lee, 1994; Dominicus et al., 2006; Verbeke and Molenberghs, 2003). However, this result depends on the assumption of independent and identically distributed (i.i.d.) data (Crainiceanu, 2008; Crainiceanu and Ruppert, 2004a, b, c), which is violated in the heritability problem. It has been shown that 0 values occur at a rate greater than 50%, producing conservative inferences (Blangero et al., 2013; Crainiceanu and Ruppert, 2004a; Shephard, 1993; Shephard and Harvey, 1990).

As with most statistical models, the quantitative genetic models used here are based on an assumption of multivariate Gaussianity, and this assumption is the basis of the estimation and test procedures described above. However, the heritability test statistic's null distribution may be inaccurate even when Gaussianity is perfectly satisfied, due to the limitations of the 50:50 $\chi^2$ result just mentioned. Further, for neuroimaging spatial statistics, like family-wise error (FWE) corrected inference with either voxel- or cluster-wise inference, the relevant parametric null distributions are intractable. While random field theory (Worsley et al., 1992; Friston et al., 1994; Nichols and Hayasaka, 2003) results exist for $\chi^2$ images (Cao, 1999), they are not directly applicable here as the test statistic image cannot be expressed as a linear combination of component error fields.

Hence, there is a compelling need for alternative inference procedures that make fewer assumptions. Permutation tests are a type of nonparametric test that can provide exact control — or approximately exact when there are nuisance variables — over false positive rates. These tests depend only on minimal assumptions, namely, that under the null hypothesis the data is exchangeable, that is, that the joint distribution of the data remains unaltered after permutation (Nichols and Holmes, 2002; Winkler et al., 2014).

There is relatively little work on permutation tests for variance component inference. The typical application of variance components models is not in quantitative genetics, but in hierarchical linear models where observational units are nested in clusters, such repeated measures designs. Of the few permutation methods proposed in this setting, they all permute the residuals (after removing the covariate effects) between and within clusters while fixing the model structure. While these procedures use different test statistics, e.g. Fitzmaurice and Lipsitz (2007) used the LRT as the statistic, while Lee and Braun (2012) used the sample variance of estimated random effect, they generally require iterative optimization of the likelihood function, and thus as permutation procedures they are yet more computationally demanding.

Samuh et al. (2012) presented a fast permutation test, though it is only applicable to the random intercept model. And recently Drikvandi et al. (2013) introduced a fast permutation test based on the variance least square estimator, which in essence fits a regression model to squared residuals. However, this approach is not based on maximum likelihood, and is only intended for a standard repeated measures model, where independent subjects are recorded multiple times, not multiple dependent subjects as in a pedigree study.

Our group presented a method to accelerate maximum likelihood estimation by applying an orthonormal data transformation that diagonalizes the phenotypic covariance, transforming a correlated heritability model into an independent but heterogeneous variance model (Blangero et al., 2013). However, this advance doesn't eliminate iterative optimization nor possible convergence problems.

In the present work, we expanded upon this work to derive approximate, non-iterative estimates and test statistics based on the first iteration of Newton's method. These procedures can be constructed with an auxiliary model based on regressing squared residuals on the kinship matrix eigenvalues. Then the Wald and score hypothesis tests can then be seen as generalized and ordinary explained sum of squares of the auxiliary model. In addition, as the null hypothesis of no heritability corresponds to homogeneous variance of the transformed phenotype, we draw from the statistical literature on tests of heteroscedasticity for a new and completely different test for heritability detection. We develop permutation test procedures for each of these methods, thus providing FWE-corrected voxel- and cluster-wise inferences.

The remainder of this paper is organized as follows. In the next section we detail the statistical model used and describe each of our proposed methods. The simulation framework used to evaluate the methods, and the real data analysis used for illustration are described in the Evaluation section. We then present and interpret results, and offer concluding remarks.

## Theory

In this section we detail the statistical models used, introduce our fast heritability estimators and tests, and then propose several permutation strategies for these tests.

### *Original and eigensimplified polygenic models*

At each voxel/element, a polygenic model for the phenotype $Y$ measured on $N$ individuals can be written as

$$Y = X\beta + g + \epsilon \tag{1}$$

where $X$ is an $N \times p$ matrix consisting of an intercept and covariates, like ageand sex; $\beta$ is the $p$-vector of regression coefficients; $g$ is the $N$-vector of latent (unobserved) additive genetic effect; and $\epsilon$ is the $N$-vector of residual errors. In this study we consider the most common variance components model, with only additive and unique environmental components.

The trait covariance, $Var(Y) = Var(g + \epsilon) = \Sigma$ can be written as

$$\Sigma = 2\sigma_A^2\Phi + \sigma_E^2 I, \tag{2}$$

where $\Phi$ is the kinship matrix; $\sigma_A^2$ and $\sigma_E^2$ are the additive genetic and the environmental variance components, respectively; and $I$ is the identity matrix. The kinship matrix is comprised of kinship coefficients, half the expected proportion of genetic material shared between each pair of individuals (Lange, 2003).

The narrow sense heritability is

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}. \tag{3}$$

Maximum likelihood is used for parameter estimation with the assumption that the data follows a multivariate normal distribution. The log likelihood for the untransformed model (Eqs. (1) & (2)) is

$$\ell(\beta, \Sigma; Y, X) = -\frac{1}{2}Nlog(2\pi) - \frac{1}{2}log(|\Sigma|) - \frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta). \tag{4}$$

For large datasets with arbitrary family structure, the computational burden of evaluating of the likelihood can be substantial. In particular, a quadratic form of the inverse covariance, $\Sigma^{-1}$, must be computed, along with the determinant of $\Sigma$. We take the approach of Blangero et al. (2013), who proposed an orthogonal transformation based on the eigenvectors of the kinship matrix, thus diagonalizing the covariance and simplifying the computation of the likelihood (Eq. (4)).

The eigensimplified polygenic model is obtained by transforming the data and model with a matrix $S$, the matrix of eigenvectors of $\Phi$ which are the same as the eigenvectors of $\Sigma$, Eq. (2). Applying this transformation to Eq. (1) gives the transformed model

$$S'Y = S'X\beta + S'g + S'\varepsilon$$